NCBI News

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Department of Health and Human Services

Volume 16, Issue 1
January 2008

The Database of Genotypes and Phenotypes(dbGaP)

WGA Study	Variables	Participants	Embargo Release Date	Analyses
Framingham SNP Health Association Resource (SHARe)	13,183	15,876	October 1, 2008	yes
International ADHD Genetics Project	438	2,835	March 26, 2008	no*
Search for Susceptibility Genes for Diabetic Nephropathy in Type 1 Diabetes	-	1,835	July 16, 2008	no*
Major Depression: Stage 1 Genome-wide Association in Population-Based Samples	-	3,786	August 16, 2008	no*
National Eye Institute (NEI) Age-Related Eye Disease Study (AREDS)	174	600	June 11, 2007	yes
National Institute of Neurological Disorders and Stroke (NINDS) Parkinson's Disease	43	1,283	October 12, 2007	yes

Table 1. Studies with whole genome association (WGA) data in dbGaP. Investigators who have submitted data to dbGaP retain the exclusive publication rights for a period of approximately 9-12 months after the data are released in dbGaP. The embargo release date is the date that exclusive publication rights expire.

Understanding the relationship between the genetic makeup of an organism (genotype) and measurable traits and responses (phenotype) is essential for the accurate diagnosis, treatment, and prevention of human disease. A growing number of large-scale and long-term studies involving many individuals are producing data sets of genomic characteristics associated with complex traits and outcomes. The Genotype and Phenotype Database (dbGaP) has been established at the NCBI to archive, distribute, and support the submission of data that correlate genomic characteristics with observable traits (Mailman, MD, et al. 2007. The NCBI dbGaP database of genotypes and phenotypes. Nature Genetics 39(10):1181-6 PMID: 17898773). The predominant sources of data in dbGaP are whole genome association (WGA) studies. WGA data are contributed by a number of projects including the Genetic Association Information Network (GAIN), the Framingham SNP Health Association Resource (SHARe), and research centers at the National Institutes of Health (NIH) (Table 1). Other data may be from medical sequencing, molecular diagnostic assays, and surveys of association between genotype and non-clinical traits. More information on WGA studies and the NIH-funded programs that support them is available on the WGA homepage:

www.ncbi.nlm.nih.gov/WGA

Data in dbGaP

NCBI organizes dbGaP data into four different data types: Studies, Study Documents, Phenotypic Variables, and Genotype-Phenotype Analyses. Completed studies have associated Documents and Variables and some have pre-computed analyses. A summary of the completed WGA studies is shown in Table 1.

Access to dbGaP Data

NCBI assigns unique identifiers (accessions) to the data in dbGaP, and researchers can search dbGaP data as a part of the NCBI Entrez system.

www.ncbi.nlm.nih.gov/sites/entrez?db=gap

The dbGaP Homepage also has a convenient browser that allows direct access to the Studies, Variables, and Documents and Analyses.

Open-access data may also be downloaded from the dbGaP ftp site.

ftp.ncbi.nlm.nih.gov/dbgap

continued on page 6

In this issue

NCBI News

NCBI News is distributed four times a year. We welcome communication from users of NCBI databases and software and invite suggestions for articles in future issues. Send correspondence to *NCBI News* at the address below. To subscribe to NCBI News, send your name and address to either the street or E-mail address below.

NCBI News
National Library of Medicine
Bldg. 38A, Room 3S-308
8600 Rockville Pike
Bethesda, MD 20894
Phone: (301) 496-2475
Fax: (301) 480-9241
E-mail: info@ncbi.nlm.nih.gov

Editors
Dennis Benson
David Wheeler

Contributors
Susan Dombrowski
Tao Tao

Writers
Peter Cooper
Rana Morris
Eric Sayers
Robert Yates

Editing and Production
Robert Yates

Print & Web Design
Robert Yates

In 1988, Congress established the National Center for Biotechnology Information as part of the National Library of Medicine; its charge is to create information systems for molecular biology and genetics data and perform research in computational molecular biology.

The contents of this newsletter may be reprinted without permission. The mention of trade names, commercial products, or organizations does not imply endorsement by NCBI, NIH, or the U.S. Government.

NIH Publication No. 08-3272

ISSN 1060-8788
ISSN 1098-8408 (Online Version)

Web BLAST Interface Re-designed

The NCBI BLAST Web services have a new organization with a simplified interface that provides easier access to important options. The redesign also offers several new features that include a more powerful taxonomic limit, automatic adjustment of search parameters for short sequences, easy tracking and access to recent searches, and the ability to store search strategies.

The New BLAST Homepage

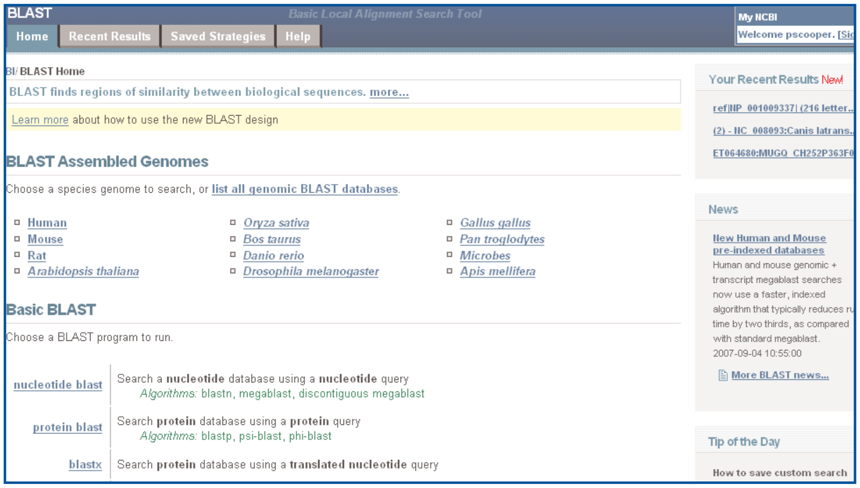
The re-designed BLAST homepage divides the online BLAST services among three logical sections (Figure 1): **Assembled Genomes**, **Basic BLAST** and **Specialized BLAST**. These sections reflect the nature of the sequence databases searched.

The **Assembled Genomes** services section of the BLAST homepage provides searches against databases of assembled genomic sequence, as well as transcripts and proteins associated with the organism and the annotated genome. All of these pages allow nucleotide, protein and translating searches from the same search submission page. For genomes in the Map Viewer, the BLAST results can be displayed on the sequence map for the organism.

The **Specialized BLAST** services feature a variety of databases including sequences from the nucleotide polymorphism database (dbSNP), reporter sequences used in gene expression experiments (GEO), germline and protein sequences of mouse and human immunoglobulins, the trace archive, and the popular BLAST 2 sequences utility. Conserved domain database searches and conserved domain architecture searches are also available here independent of ordinary protein BLAST searches.

The **Basic BLAST** section of the homepage provides access to the standard nucleotide and protein databases. Forms are available for all five possible combinations of query sequence and database: nucleotide query - nucleotide database, protein query - protein database, nucleotide

continued on page 8



BI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

Learn more about how to use the new BLAST design

BLAST Assembled Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query

Your Recent Results New!

ref|NP_001009337| (216 letter...
(2) - NC_008093:Canis latrans...
ET064680:MUGQ_CH252P363F0

News

New Human and Mouse pre-indexed databases
Human and mouse genomic + transcript megablast searches now use a faster, indexed algorithm that typically reduces ru time by two thirds, as compared with standard megablast.
2007-09-04 10:55:00

More BLAST news...

Tip of the Day

How to save custom search

Figure 1. The new BLAST homepage with access to Assembled Genome forms, universal Basic BLAST forms, and specialized BLAST services. Tabs at the top of the page provide access to Recent Results - searches less than 36 hours old, Saved Search Strategies, and Help documentation. My NCBI account sign-in is available at the upper right on all BLAST pages and allows saving BLAST search strategies. Links to the most recent search results are also displayed. The Recent Results tab links to a table of recent search results.

New Genome Builds and Map Viewer Displays

New Map Viewer Homepage

The Map Viewer homepage has a new look and added functions. To make room for additional organisms, the page now presents expandable sections for the groups of organisms available in Map Viewer (Figure 1). As always, the Map Viewer homepage is linked directly to the NCBI homepage under the "Hotspots" list and to the Genomic BLAST section of the new BLAST homepage. The page can also be accessed directly.

www.ncbi.nlm.nih.gov/mapview

The Map Viewer now hosts 79 species: 22 multi-cellular animals, seven basal eukaryotes (protozoa), 35 green plants, and 15 fungi. The Map Viewer homepage links to the Maps, genomic BLAST, and Genome Resources pages for each of the organisms.

New Mammals in Map Viewer

The taxonomic coverage of mammalian genomes has been expanded to include the first marsupial genome, a South American opossum (*Monodelphis domestica*); an egg-laying mammal, the duck-billed platypus; and the first member of the placental order perissodactyla, the domestic horse. The platypus and the opossum genomes will provide key insights into mammalian evolution through comparative genomic studies. The horse genome is of significant veterinary interest in addition to its importance in evolutionary and comparative biology.

The horse genome build 1.1 is NCBI's assembly and annotation of the 6.8 X whole genome shotgun assembly produced by the Broad Institute. The assembly provides a total sequence length of 2.42 Gigabases and is anchored to the 31 horse autosomes and the X chromosome. The mitochondrial genome given in NCBI RefSeq **NC_001640**, derived from GenBank record **X79547**, is also displayed in the Map Viewer.

The opossum (*Monodelphis domestica*) genome build 2.1 is the assembly and annotation of the 6.5 X whole genome shotgun sequence (MonDom5) also produced by the Broad Institute. The assembly provides a total sequence length of 3.50 Gigabases and is mapped onto the 8 opossum autosomes and the X chromosome. The opossum mitochondrial genome —**NC_006299** based on GenBank **AJ508398**— is also included in the build. A total of 22,478 genes and their transcripts are placed on the opossum sequence in the Map Viewer.

The duck-billed platypus genome is less completely anchored than the other two new mammals in part because of the complex genome structure involving micro- and macro-chromosomes and multiple sex chromosomes. NCBI build 1.1 is based on the 6 X assembly of the platypus genome produced by the Washington University Genome Sequencing Center. The total sequence length is 1.84 Gigabases and is partly mapped onto 15 of the 21 platypus autosomes and four of the 10 sex chromosomes. Unlike the horse and opossum genomes, the majority of the platypus genome contigs and components are not mapped to the chromosomes. As with the horse and opossum, the platypus build includes a mitochondrial genome sequence — **NC_000891** based on **X83427**. A total of 2,945 genes and their transcripts are placed on the platypus chromosome sequence in the Map Viewer. An additional 17,781 genes and transcripts are annotated on unplaced contigs. Both mapped and unplaced contigs can be displayed in the map viewer and searched as described below for the jewel wasp genome.

The Jewel Wasp Debuts in Map Viewer

The genome of the jewel wasp, a parasitoid wasp of the Chalcidoidea is now available in Map Viewer. The jewel wasp is a member of a group of insects with complex life histories, many involving parasite-like feeding on host insects. Wasps in this group are important as potential biological control agents on host pest species and, in the case of some plant feeding members, as pests themselves. The jewel wasp genome should provide important insights into development and evolution of complex life histories and help in the production of more and better biological controls from these insects. The genome sequence is derived from a highly inbred laboratory strain (AsymCX) of the jewel wasp with well-understood genetics. NCBI build 1.1 is based on the 6.2 X assembly produced by the Human Genome Sequencing Center at the Baylor College of Medicine. The total sequence length is 239 Megabases with 10,734 annotated genes. In the current build, none of the sequences are placed on the jewel wasp chromosomes, but the unplaced sequences can be displayed in the Map Viewer and can be effectively searched by querying Map Viewer for any available maker including a gene name, gene symbol, or accession numbers. The Gene database also provides links to the Map Viewer for annotated genes. Searches with sequences on the jewel wasp genomic

continued on page 9

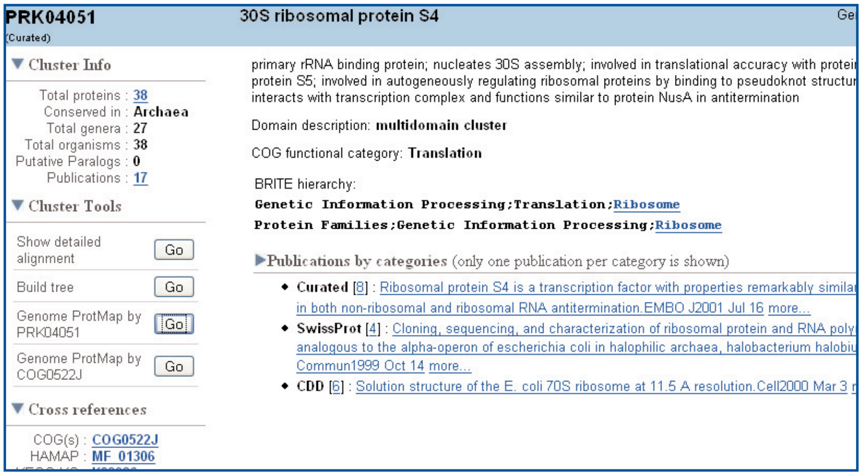
New Protein Clusters database

NCBI has released an Entrez database called Protein Clusters in an effort to facilitate rapid protein analysis and annotation. Protein clusters consists of Reference Sequence (RefSeq) proteins from complete genomes of prokaryotes, from plasmids, and from eukaryotic organelles. Protein clusters are created using a modified BLAST score that takes into account the length of the hit (alignment) on both the query and the subject. Sequences are then sorted by these modified scores, and all proteins that are contained within the top hits are clustered together. The database can be accessed from the search pulldown menu on the NCBI home page or from the URL below.

www.ncbi.nlm.nih.gov/sites/entrez?db=proteinclusters

The Protein Cluster Overview Page

A Protein Cluster record may be either curated (PRK-type identifier for prokaryotes, CHL for organelles) or non-curated (CLS for prokaryotes, CLSC for organelles). Curated clusters have consistent nomenclature and protein function descriptions that are displayed on the overview page along with annotated metabolic functions, links to NCBI Conserved Domains (CDD), and several tools for displaying protein sequence. In addition, the overview page contains links to categorized publications, including those linked to the cluster by curators and those linked to proteins from RefSeq, GenPept, SwissProt, Structure, and Conserved Domains. Figure 1 shows the overview page for PRK04051, 30S ribosomal protein S4 from Archaea. Non-curated records have similar pages, but display information that was collected automatically.



PRK04051 (Curated) 30S ribosomal protein S4

▼ Cluster Info

Total proteins :	38
Conserved in :	Archaea
Total genera :	27
Total organisms :	38
Putative Paralogs :	0
Publications :	17

▼ Cluster Tools

Show detailed alignment Go
Build tree Go
Genome ProtMap by PRK04051 Go
Genome ProtMap by COG0522J Go

▼ Cross references

COG(s) : COG0522J
HAMAP : MF_01306

primary rRNA binding protein; nucleates 30S assembly; involved in translational accuracy with protei
protein S5; involved in autogeneously regulating ribosomal proteins by binding to pseudoknot structu
interacts with transcription complex and functions similar to protein NusA in antitermination

Domain description: **multidomain cluster**

COG functional category: **Translation**

BRITE hierarchy:
Genetic Information Processing;Translation;Ribosome
Protein Families;Genetic Information Processing;Ribosome

▶**Publications by categories** (only one publication per category is shown)

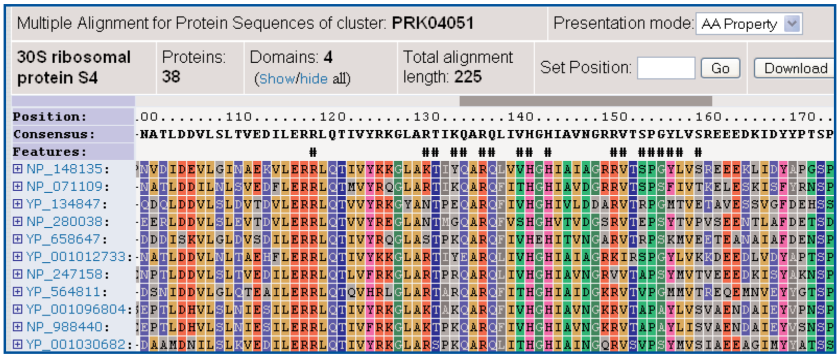
- **Curated** [8] : Ribosomal protein S4 is a transcription factor with properties remarkably simila in both non-ribosomal and ribosomal RNA antitermination.EMBO J2001 Jul 16 more...
- **SwissProt** [4] : Cloning, sequencing, and characterization of ribosomal protein and RNA poly analogous to the alpha-operon of escherichia coli in halophilic archaea, halobacterium halobiu Commun1999 Oct 14 more...
- **CDD** [6] : Solution structure of the E. coli 70S ribosome at 11.5 A resolution.Cell2000 Mar 3

Figure 1. Protein Cluster Overview Page for PRK04051.

Below the tools and links on the overview page is a table showing detailed information about each protein in the cluster. Links are provided to the organism from which the sequence was derived as well as to Entrez Protein and Gene. Any paralogs in the cluster can be easily identified by clicking the "Highlight paralogs" link above the Organism column. Upstream and downstream clusters in the local genomic neighborhood are linked, and these cluster labels are colored according to their COG functional category. Also provided are links to pre-computed BLAST results (BLink), and a graphic summary of that sequence's participation in the multiple alignment of the cluster. Dark grey bars indicate aligned regions of the given sequence in the multiple alignment, and colored bars beneath these represent the footprints of domains from CDD.

Analysis and Display Tools

Sequence Alignment. The "Show detailed alignment" tool displays a pre-computed multiple sequence alignment for the cluster. Figure 2 displays a portion of the multiple alignment for PRK04051 showing the location of the RNA binding surface. The alignment may be viewed either by amino acid properties, where residues are colored by their chemical type (charged, aromatic, hydrophobic, etc.), or by consensus, where residues are shown only if they differ from the consensus residue for the given column. The consensus sequence and a position scale are shown above the alignment for reference. In addition, the footprint of domains from CDD



Multiple Alignment for Protein Sequences of cluster: **PRK04051** Presentation mode: AA Property

30S ribosomal protein S4	Proteins: 38	Domains: 4 (Show/hide all)	Total alignment length: 225	Set Position:	Go	Download

```
Position:      .00.......110.......120.......130.......140.......150.......160.......170..
Consensus:     -NATLDDVLSLTVEDILERRLQTIVYRKGLARTIKQARQLIVHGHIAVNGRRVTSPGYLVSREEEDKIDYYPTSP
Features:                         #           ## ## ##  ## #      ## ##### #
⊞ NP_148135:    NVDIDEVLGINAEKVLERRLQTIVYKKGLAKTIYQARQLVVHGHIAIAGRRVTSPGYLVSREEEKLIDYAPGSP
⊞ NP_071109:   -NATLDDILNLSVEDFLERRLQTMVYRQGLARTIKQARQFITHGHIAVDGRRVTSPSFIVTKELESKISFYRNSP
⊞ YP_134847:   -QDQLDDVLSLDVTDVLERRLQTVVYRKGYANTPEQARQFIVHGHIVLDDARVTRPGMTVETAVESSVGFDEHSS
⊞ NP_280038:   -EERLDDVLSLEVTDVLERRLQTVVYREGLANTMGQARQFVSHGHVTVDGSRVTEPSYTVPVSEENTLAFDETSD
⊞ YP_658647:   -DDDISKVLGLDVSDILERRLQTIVYRQGLASTPKQARQFIVHEHITVNGARVTRPSKMVEETEANAIAFDENSP
⊞ YP_001012733:-NATLDDVLNLTAEHFLERRLQTIVYKKGLARTIYEARQLIVHGHIAIAGRKIRSPGYLVKKDEEDLVDYAPTSP
⊞ NP_247158:    NPTLDDVLSLTVEDILERRLQTLVFRKGLARTPRQARQLIVHGHIAVNGRVVTAPSYMVTVEEEDKISYAKNSP
⊞ YP_564811:   -DSNIDDVLGLQTEAILERRLQTQVHRLGLARTARQARQFITHGHIAIDGKRVTVPGMMVTREQEMNVEYYGTSP
⊞ YP_001096804: EPTLDHVLSLNIESILERRLQTIVYRKGLAKTAKQARQFIVHGHIAVNGKRVTAPAYLVSVAENDAIEYVPNSP
⊞ NP_988440:    EPTLDHVLSLNIESILERRLQTIVFRKGLAKTPKQARQFIVHGHIAVNGKRVTAPAYLISVAENDAIEYVSNSP
⊞ YP_001030682:-DAAMDNILSLKVEDILERRLQTIVYRKGLARSPKQARQLITHGHIAINGQRVSVPSYMVSIAEEAGIMYYATSS
```

Figure 2. Portion of the multiple sequence alignment for PRK04051, showing the region including the RNA binding surface, represented by "#" symbols in the 'Features' row.

may be shown for individual sequences or for all sequences in the alignment, and if these matching CDs are curated and contain annotated features, then their locations will be marked by '#' symbols in the Features row above the alignment. The alignment may also be downloaded as a multiple FASTA text file.

Phylogenetic Trees. The "Build tree" tool displays an interactive phylogenetic tree of the sequences in the cluster. The tool supports two distance methods: neighbor-joining and fast minimum evolution, with multiple distance measures available for each method. Any branch of the tree can be expanded, collapsed, or squeezed, or it can be used to re-root the entire tree. Users can re-root the tree by clicking on any node, which will then become the new root. The tool also supports automatic collapsing by taxonomic rank, allowing, for example, only nodes of particular rank or higher to remain expanded.

ProtMap. The ProtMap tool provides a graphical view of the local genomic neighborhood for genes represented by sequences in a Protein Cluster. These genes are highlighted yellow in the center of the display, while the adjacent genes are colored by their COG functional category. The view is interactive, with each gene bar providing links to Entrez Protein, Gene, or a new ProtMap view with the given gene as the reference. Figure 3 displays a portion of the region near PRK04051 for three families of Crenarchaeota. The S4 gene is immediately downstream of S13 in all genera in these families, but in the Sulfolobaceae and Thermofilaceae, S4 is immediately upstream of S11, while in the Thermoproteaceae S4 is upstream of asparaginase 2.

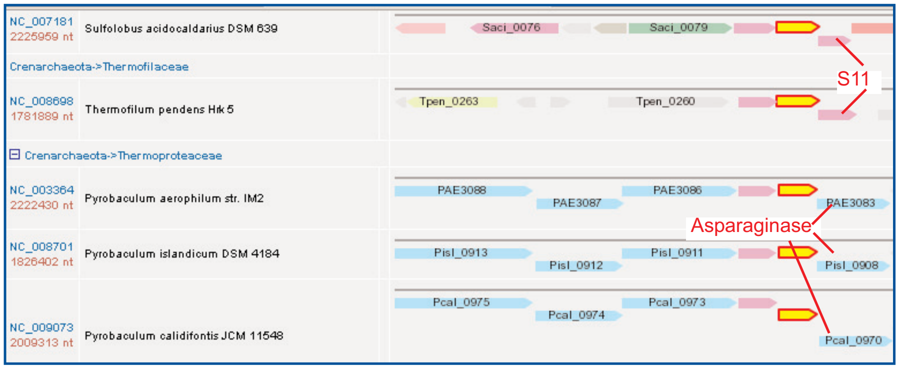


Figure 3. Portion of the ProtMap for PRK04051. The S4 gene is highlighted in yellow. The pink genes upstream and downstream of S4 are S13 and S11, respectively in the Thermofilaceae and Sulfolobaceae (top of image). The genes downstream of S4 in the Thermoproteaceae represent asparaginase 2.

Finding Protein Clusters

Entrez queries. The Protein Clusters database supports all standard Entrez search and indexing functions, including the Limits and Preview/Index tab that assist users in building advanced queries. For example, users may limit queries by a gene name, EBI HAMAP identifier, KEGG orthology identifier, organism, protein accession number, sequence length, the number of proteins in a cluster, or the title of the cluster. A detailed description of the indexing fields in Protein Clusters is provided at

www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helpcluster.chapter.helpcluster

Entrez links. Protein Clusters have reciprocal links to PubMed abstracts, protein sequences, genomic sequences (in both Entrez CoreNucleotide and Genome), and genes. Users may therefore find Protein Cluster records by first searching for a gene of interest, a PubMed abstract, or a genome, and then following the link to Protein Clusters.

Concise Protein BLAST. NCBI has created a specialized BLAST service to support Protein Clusters. The database for this service includes, for each cluster, one representative protein sequence from each genus in the cluster, thereby reducing sequence redundancy and improving search performance. For completeness, the database includes sequences from both curated and non-curated clusters, as well as sequences that were not placed in a Protein Cluster. Concise Protein BLAST supports both blastp (protein queries) and blastx (nucleotide queries) and allows users to adjust the word size, scoring matrix, expect-value cutoffs, and gap penalties. The service presents the results in a table showing the genus and accession of each hit along with its length, locus tag, Protein Cluster accession, BLAST statistics, and links to views of the pairwise alignment.

CD-Search. NCBI has calculated position-specific scoring matrices (PSSMs) from the curated Protein Cluster alignments, and has entered these records into Entrez CDD. Beginning with CDD release v2.12, these PSSMs form part of the default search database for the CD-Search service, allowing Protein Cluster records to be matched to sequences using RPS-BLAST (CD-Search). Moreover, these PSSMs enable users to collect other related protein sequences that are not part of the curated alignment, as well as to view and manipulate the alignments with other NCBI software such as Cn3D and CD-Tree. ¤

dbGaP

continued from page 1

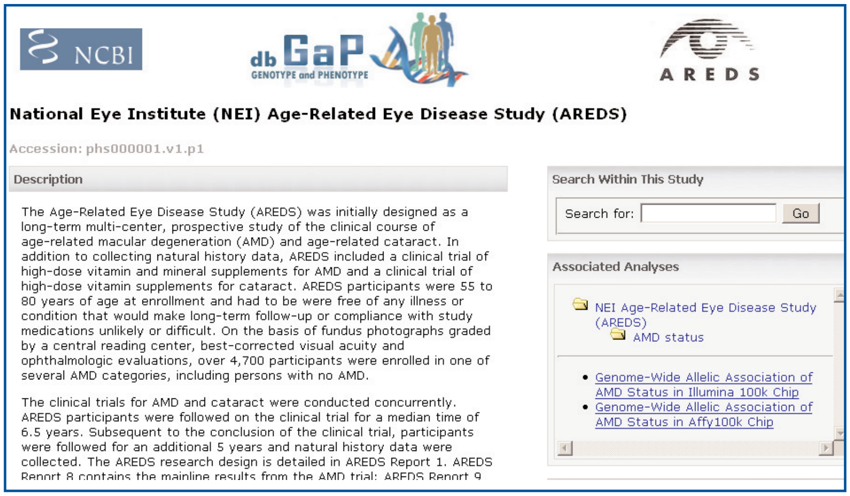
Data and Privacy Considerations

Studies may have open-access data and analyses as well as controlled-access individual level data available. To protect the identity of the individuals involved in the studies, NCBI only accepts data with anonymous identifiers. Nevertheless, some of the individual-level data in certain studies may be detailed enough that privacy could be compromised. Therefore, access to individual level data is controlled and only available to approved researchers. Approval is granted by an NIH Data Access Committee (DAC). Researchers wishing access to controlled data must submit a Data Use Certification (DUC) to the appropriate NIH DAC for approval. Details on access to controlled data and the application process are available from the dbGaP pages.

dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?login=&page=login

Displaying Data and Analyses in dbGaP

Figure 1 shows the summary page record in the dbGAP Entrez service for the National Eye Institute (NEI) Age-Related Eye Disease Study (AREDS). The summary provides access to associated documents, variables, and analyses, as well as links to the public data on the ftp site and a request for access to individual level data. In addition there are links to publications in PubMed associated with the study and a list of authorized individual level data requests. In the AREDS data the genome wide association variable Age-Related Macular Degeneration (AMD) Status has associated analyses that can be displayed in the dbGaP Genome Viewer and Chromosome Browser (Figure 2 shown on page 7.) The Genome Viewer quickly shows regions of the genome that contain phenotype-associated alleles. Linking to the Chromosome Browser provides detailed information about the alleles from dbSNP and precise locations on the NCBI Map Viewer with links to Entrez Gene and Entrez Nucleotide. In the case of AMD status, there are strongly associated polymorphisms on chromosome 1 that fall within the Regulator of Complement Activation gene cluster. Several fall within the Complement Factor H (CFH) gene. These AREDs data have been used to firmly establish a link between CFH polymorphisms and AMD, a major cause of blindness in the elderly (Klein, R.J. et al. 2005. Complement factor H in age-related macular degeneration. *Science*, 308(5720):385-9 PMID: 15761122).



National Eye Institute (NEI) Age-Related Eye Disease Study (AREDS)

Accession: phs000001.v1.p1

Description

The Age-Related Eye Disease Study (AREDS) was initially designed as a long-term multi-center, prospective study of the clinical course of age-related macular degeneration (AMD) and age-related cataract. In addition to collecting natural history data, AREDS included a clinical trial of high-dose vitamin and mineral supplements for AMD and a clinical trial of high-dose vitamin supplements for cataract. AREDS participants were 55 to 80 years of age at enrollment and had to be were free of any illness or condition that would make long-term follow-up or compliance with study medications unlikely or difficult. On the basis of fundus photographs graded by a central reading center, best-corrected visual acuity and ophthalmologic evaluations, over 4,700 participants were enrolled in one of several AMD categories, including persons with no AMD.

The clinical trials for AMD and cataract were conducted concurrently. AREDS participants were followed on the clinical trial for a median time of 6.5 years. Subsequent to the conclusion of the clinical trial, participants were followed for an additional 5 years and natural history data were collected. The AREDS research design is detailed in AREDS Report 1. AREDS Report 8 contains the mainline results from the AMD trial; AREDS Report 9

NEI Age-Related Eye Disease Study (AREDS)
AMD status

- Genome-Wide Allelic Association of AMD Status in Illumina 100k Chip
- Genome-Wide Allelic Association of AMD Status in Affy100k Chip

Figure 1. A portion of the dbGaP summary page for the Age-Related Eye Disease Study (AREDS). The phenotypic variable AMD status has associated genome wide association analyses in dbGAP.

Summary

NCBI's Database of Genotypes and Phenotypes fills a critical and growing need as a public repository for phenotype and genotype data and associated analyses, and provides a uniform representation for these data. dbGaP is prepared for the future growth and change in scope of these important data through a flexible database structure that can handle a wide range of genotype and phenotype data. The database is prepared to accept data beyond the human GWAS high density microarray genotype data as explored here, including data from new technologies, non-clinical human and model organism data,

continued on page 7

GenBank Release 164

GenBank Release 164 (February 2008) contains over 82 million sequence entries totaling more than 85 billion base pairs. Release 165 is scheduled for April 2008. GenBank is accessible via the Entrez search and retrieval system. The flatfile and ASN.1 versions of the Release are found in the "genbank" and "ncbi-asn1" directories respectively at:

ftp.ncbi.nih.gov

Uncompressed, the Release 164 flatfiles are 321 Gigabytes and the ASN.1 version is about 295 Gigabytes. The data can also be downloaded at a mirror site:

bio-mirror.net/biomirror/genbank

dbGaP
continued from page 6

even agricultural genotypes and phenotypes.

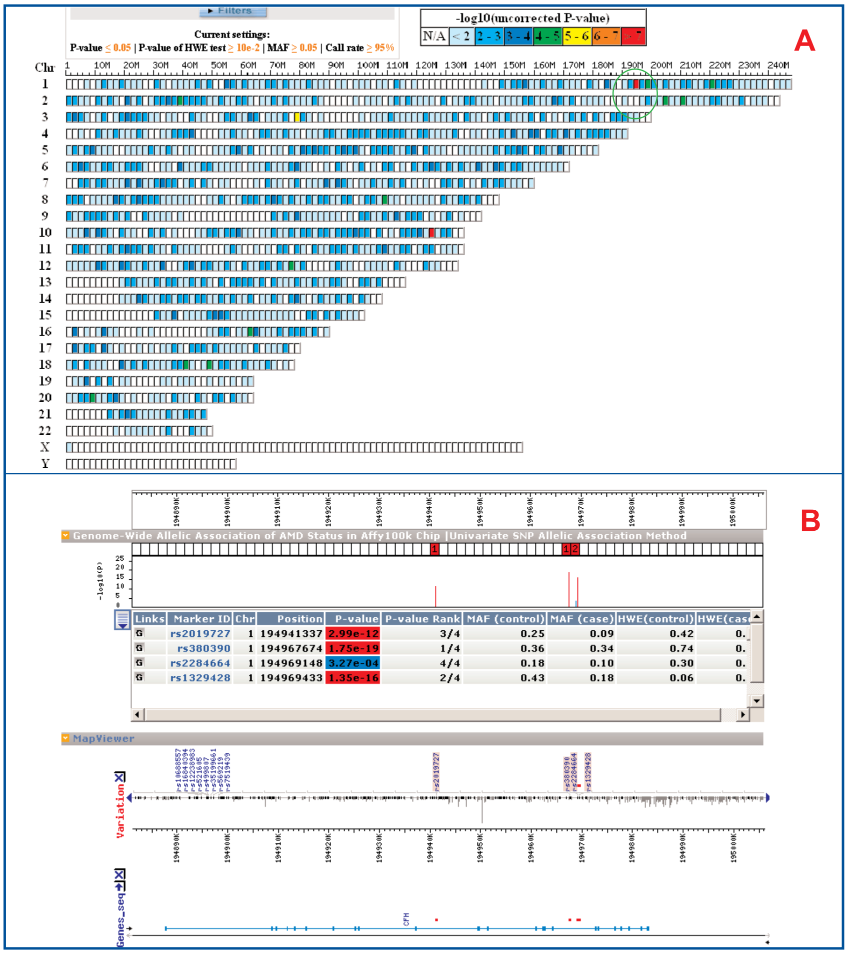
Questions about access to individual-level data or dbGaP submissions should be sent to the dbGaP helpdesk

dbgap-help@ncbi.nlm.nih.gov

Questions about searching public data at NCBI should be sent to the NCBI Service Desk.

info@ncbi.nlm.nih.gov

—SD



Current settings:
P-value ≤ 0.05 | P-value of HWE test ≥ 10e-2 | MAF ≥ 0.05 | Call rate ≥ 95%

Links	Marker ID	Chr	Position	P-value	P-value Rank	MAF (control)	MAF (case)	HWE(control)	HWE(case
G	rs2019727	1	194941337	2.99e-12	3/4	0.25	0.09	0.42	0.
G	rs380390	1	194967674	1.75e-19	1/4	0.36	0.34	0.74	0.
G	rs2284664	1	194969148	3.27e-04	4/4	0.18	0.10	0.30	0.
G	rs1329428	1	194969433	1.35e-16	2/4	0.43	0.18	0.06	0.

Figure 2. Genome wide analysis of polymorphisms associated with AMD status in the AREDS data. Polymorphisms were genotyped using the Affymetrix GeneChip Human Mapping 100K Set. A. The dbGaP Genome Viewer showing the location of the sequence polymorphisms associated with Age-Related Macular Degeneration (AMD) and their degree of association (uncorrected P-value). B. Detailed view of the highly associated region on chromosome 1 in the dbGaP Chromosome Browser showing the locations of associated SNPs on the Complement Factor H gene.

RefSeq Release 27

RefSeq Release 27 is now available by anonymous FTP at:

ftp.ncbi.nih.gov/refseq/release

Release 27 includes genomic, transcript, and protein sequences available as of January 6, 2008, from 4,926 organisms. The number of RefSeq accessions in Release 27 and their combined lengths is given in the shaded box.

	# of Accessions	# of Basepairs/Residues
Genomic	1,387,692	99,006,517,014
RNA	1,211,414	2,053,035,099
Protein	4,426,609	1,556,356,987

RefSeq releases are posted every two months, and the next release is scheduled for March, 2008. Release notes documenting the scope and content of the database are provided at:

ftp.ncbi.nih.gov/refseq/release/release-notes

For more information, visit the NCBI RefSeq Web Site at:

www.ncbi.nih.gov/RefSeq

Web BLAST Interface

continued from page 2

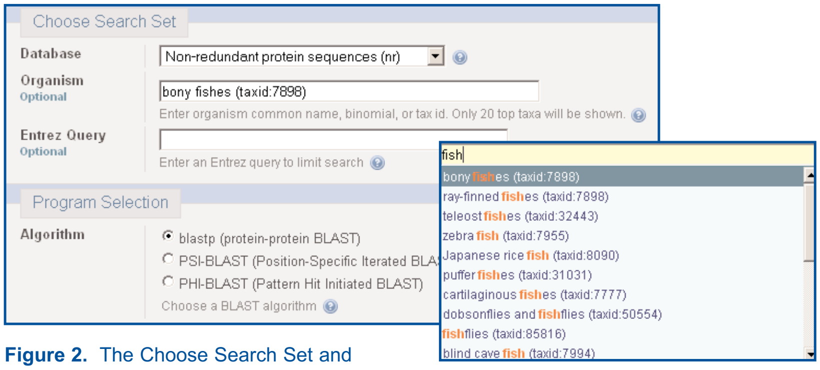
translation - protein database, protein query - nucleotide translation, nucleotide translation - nucleotide translation. The nucleotide BLAST form allows the selection of megablast, discontiguous megablast or the traditional blastn search. The protein BLAST form allows the selection of traditional blastp as well as position specific searches (PSI - and PHI- BLAST).

New Features on the Basic BLAST Submission Forms

The new submission forms are simplified, showing only the most commonly used BLAST options by default. These include database choices and the ability to restrict searches to database subsets using an organism or an Entrez query limit and program selection (algorithm) for the Basic Nucleotide and Protein forms.

Organism Limits

An important new feature of the basic forms is the re-designed organism limit box. Any valid NCBI taxonomic group can be typed in the box to restrict the database to that taxonomic subset. The organism box has an auto complete feature that shows taxonomic matches to whatever is typed, even a partial word. The auto complete aspect is very helpful when the exact taxonomic name used by the NCBI is uncertain or part of the name is known. For example, the word "fish" finds matches in the taxonomy database to several likely taxa: bony fishes, teleost fishes, ray-finned fishes, cartilaginous fishes (Figure 2). These new organism limits use pre-compiled identifier lists and are faster than using Entrez organism queries to restrict the search.



Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism (Optional): bony fishes (taxid:7898)

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Entrez Query (Optional):

Enter an Entrez query to limit search

fish

- bony fishes (taxid:7898)
- ray-finned fishes (taxid:7898)
- teleost fishes (taxid:32443)
- zebra fish (taxid:7955)
- Japanese rice fish (taxid:8090)
- puffer fishes (taxid:31031)
- cartilaginous fishes (taxid:7777)
- dobsonflies and fishflies (taxid:50554)
- fishflies (taxid:85816)
- blind cave fish (taxid:7994)

Program Selection

Algorithm:

- ● blastp (protein-protein BLAST)
- ○ PSI-BLAST (Position-Specific Iterated BLA
- ○ PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

Figure 2. The Choose Search Set and Program Selection portion of the universal protein BLAST form. The organism auto complete offers several matches from the NCBI taxonomy database for the word "fish."

Universal Forms

There are now universal nucleotide and protein basic forms that access the various BLAST programs. For basic nucleotide searches, the formerly separate megablast, discontiguous megablast and blastn forms are now combined in a single form. The program choices are available as radio buttons in the Program selection section of the form. The fastest algorithm, Megablast, is selected by default. Discontiguous megablast and blastn can be selected for more sensitive searches. For protein searches the Program selection allows the choice of blastp, or position specific searches (PSI and PHI-BLAST). When the PHI-BLAST radio button is selected a box appears that allows the seed pattern to be entered.

Algorithm Parameters and Short Sequences

The more advanced settings are available through the expandable "Algorithm Parameters" section at the bottom of the basic forms. Many of these advanced options adjust automatically to those most appropriate to the choice of database and algorithm in the upper portion of the form. For instance, when different nucleotide algorithms are chosen, the word size, and match mismatch penalty are changed appropriately. The settings also adjust automatically when a short nucleotide or protein sequence is entered in the query box eliminating the need for previously separate pages optimized for short nucleotide and protein queries.

Recent Results and Saved Search Strategies

There are now tabs on the BLAST pages that provide access to Recent Results and, through a My NCBI account, "Saved Search Stategies". NCBI Web BLAST results are stored on our servers and can be retrieved and reformatted for up to 36 hours through the Request Identifier (RID) assigned to each search. A less complex format of the RIDs makes them easier to copy and paste for saving if desired. However, the "Recent Results" tab makes local saving of RIDs unnecessary since the page gives the history of recent BLAST searches in the form of a table of recent results with a link to each unexpired results set. Each set of results is identified by its submission time, RID, search title, query length, and database searched (Figure 3 on page 9). The results can be sorted by each of these identifiers. Search strategies including query sequence, database, organism or Entrez limits and any algorithm parameters can be saved from the "Recent Results" page. This is ideal for repeating

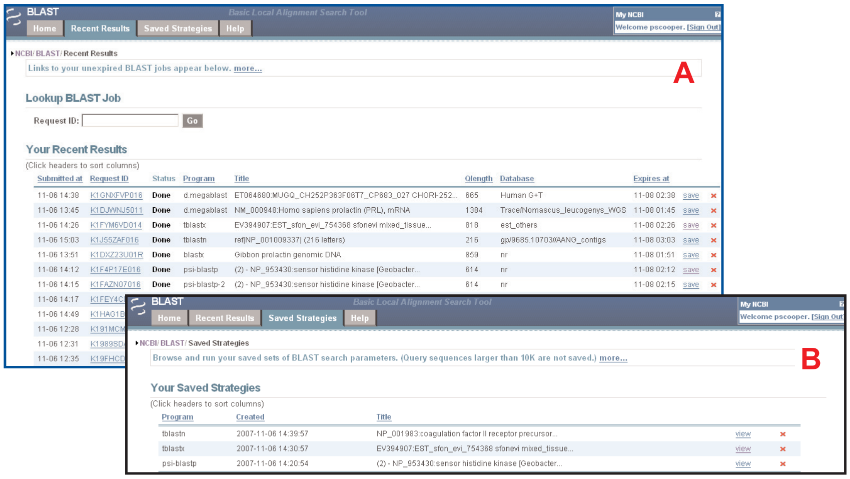
continued on page 9

Web BLAST Interface
continued from page 8

searches with the same settings to find new matches as the databases grow. Saving search strategies is made possible through My NCBI. Registering for a My NCBI account is free and available through the My NCBI link that is now at the top of all BLAST Web pages. In addition to saving BLAST search strategies, My NCBI provides a number of useful ways to customize access to NCBI services. See: Have it your way with MyNCBI!. *NCBI News.* 2005 May; 14(1):1, for more information about My NCBI.

Formatting page changes

The BLAST formatting page no longer appears by default with searches from the Basic BLAST pages. Clicking on the "Reformat these Results" link at the top of any results page invokes the formatting page. In addition to the reformatting options that were previously available, the new organism filter is available with the auto complete feature; and, for position specific searches (PSI / PHI-BLAST) the new portable Scoremat format is available. This feature is described in more detail in the BLAST Lab in this issue.


BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
Welcome pscooper. [Sign Out]
A
Go
save
BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
Welcome pscooper. [Sign Out]
B
view

NCBI/ BLAST/ Recent Results

Links to your unexpired BLAST jobs appear below. more...

Lookup BLAST Job

Request ID:

Your Recent Results

(Click headers to sort columns)

Submitted at	Request ID	Status	Program	Title	Qlength	Database	Expires at		
11-06 14:38	K1GNXFVP016	Done	d.megablast	ET064680:MUGQ_CH252P363F06T7_CP683_027 CHORI-252...	665	Human G+T	11-08 02:38	save	✕
11-06 13:45	K1DJWNJ5011	Done	d.megablast	NM_000948:Homo sapiens prolactin (PRL), mRNA	1384	Trace/Nomascus_leucogenys_WGS	11-08 01:45	save	✕
11-06 14:26	K1FYM6VD014	Done	tblastx	EV394907:EST_sfon_evi_754368 sfonevi mixed_tissue...	818	est_others	11-08 02:26	save	✕
11-06 15:03	K1J55ZAF016	Done	tblastn	ref|NP_001009337| (216 letters)	216	gp/9685.10703//AANG_contigs	11-08 03:03	save	✕
11-06 13:51	K1DXZ23U01R	Done	blastx	Gibbon prolactin genomic DNA	859	nr	11-08 01:51	save	✕
11-06 14:12	K1F4P17E016	Done	psi-blastp	(2) - NP_953430:sensor histidine kinase [Geobacter...	614	nr	11-08 02:12	save	✕
11-06 14:15	K1FAZN07016	Done	psi-blastp-2	(2) - NP_953430:sensor histidine kinase [Geobacter...	614	nr	11-08 02:15	save	✕
11-06 14:17	K1FEY4C								
11-06 14:49	K1HAG1B								
11-06 12:28	K191MCM								
11-06 12:31	K1989SD								
11-06 12:35	K19FHCD								

NCBI/ BLAST/ Saved Strategies

Browse and run your saved sets of BLAST search parameters. (Query sequences larger than 10K are not saved.) more...

Your Saved Strategies

(Click headers to sort columns)

Program	Created	Title		
tblastn	2007-11-06 14:39:57	NP_001983:coagulation factor II receptor precursor...	view	✕
tblastx	2007-11-06 14:30:57	EV394907:EST_sfon_evi_754368 sfonevi mixed_tissue...	view	✕
psi-blastp	2007-11-06 14:20:54	(2) - NP_953430:sensor histidine kinase [Geobacter...	view	✕

Figure 3. The Recent Results (A) and Saved Strategies (B) pages. Recent Results are provided as a sortable table of results that can be retrieved and reformatted up to the expiration times. Saved search strategies (B) are stored through a MyNCBI account and can be re-run at anytime to update search results.

The new Web BLAST interface and associated new features should provide researchers with a more powerful and flexible tool for sequence similarity searching. Questions and comments about the new interface should be directed to the BLAST help-desk:

blast-help@ncbi.nlm.nih.gov

¤

Genome Builds and Maps
continued from page 3

BLAST page also provide direct links to the Map Viewer from the results.

For all animal genomes mentioned, sequence maps available within Map Viewer include the NCBI contigs, the WGS sequences and the location of genes, Reference Sequence transcripts, STSs, ESTs, UniGene clusters and Gnomon predicted gene models.

Phytoplankter Genome

The genome of *Ostreococcus lucimarinus*, the first complete green alga genome, is now available in Map Viewer. The free-living *Ostreococcus lucimarinus* cells are among the smallest eukaryotic cells. These tiny algae belong to the smallest size fraction of marine phytoplankton, the picoplankton, that are among the most important primary producers on earth. Picoplankton significantly affect global biogeochemical processes and food webs and are surprisingly numerous and diverse. The *Ostreococcus* genome will provide insights into cell physiology and production in this important group. It will also shed light on mechanisms of speciation in the picoplankton, the evolution of small cell size as well as the evolution and origin of higher plants. The genome at the NCBI is the finished and annotated genome submitted by the Joint Genome Institute. The submitted sequence is 13.2 Megabases comprising the 21 chromosomes with 7,603 annotated genes and gene models.

Leishmania Genomes

Leishmaniasis, a parasitic disease caused by various species in the protozoan genus *Leishmania*, is an important health problem in much of the tropical parts of the world, causing disfiguring cutaneous lesions and more serious complications. NCBI now has the genomes of two of the human pathogenic species of *Leishmania* (*L. braziliensis and L. infantum*). Both genomes will be important in further understanding the biology of these parasites and will help in developing new strategies to combat this disease. The *braziliensis* and *infantum* genomes are comprised of completely annotated chromosome records with associated genes and model transcripts. The *braziliensis* genome was submitted by the Wellcome Trust Sanger Institute and has 31.4 Megabases comprising the 35 chromosomes with 8,129 annotated genes and gene models. The *infan-*

continued on page 10

Genome Builds and Maps

continued from page 9

tum genome submitted by a consortium that included The Wellcome Trust Sanger Institute, Imperial College, and University of Glasgow has 32.1 Megabases comprising the 36 chromosomes with 8,186 annotated genes and gene models.

For the *Ostreococcus* and *Leishmania* genomes, the contig, gene, and transcript maps are available within Map Viewer.

Updated genome builds

Mouse Build 37.1

NCBI has produced a new assembly and annotation of the mouse genome, Build 37.1. The current NCBI build shows 32,533 genes placed on the reference assembly of the C57BL/6J mouse strain. This is a mixed whole genome shotgun (WGS) and BAC clone assembly and contains a greater proportion of finished BAC clone sequence than the previous build. Several alternate assemblies are also available in the Map Viewer including a mixed-strain WGS assembly from Celera, the MGSCv3 assembly, and several BAC-based partial assemblies from 16 other mouse strains.

Zebrafish Build 2.1 (Zv6)

The NCBI now hosts the sixth assembly of the zebrafish genome (Zv6) produced by the Wellcome Trust Sanger Institute. This is a 6.5-7X mixed WGS and BAC clone assembly tied to the tiling path provided by the March 12th, 2006 fingerprint contig map. This map is comprised of more than 1 Gigabase of DNA sequence from predominantly finished BAC clones. Gaps in the scaffold are filled with WGS sequence. The NCBI zebrafish build 2.1 includes the 1.5 Gigabase Zv6 assembly and a complete mitochondrial genome of a strain ABC zebrafish (**NC_002333** based on **AC024175**). The NCBI build 2.1 annotation places 34,362 genes and their transcripts on the 25 zebrafish chromosomes.

Arabidopsis thaliana Build 7.0

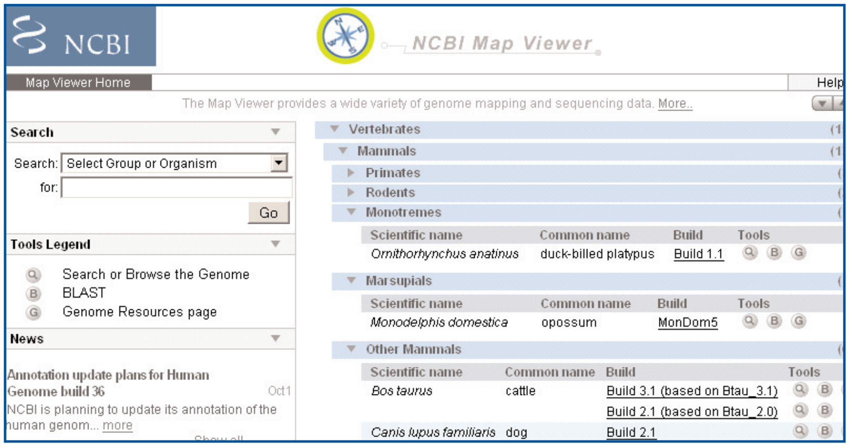
The *Arabidopsis* Information Resource release 7 (TAIR7) of the *Arabidopsis* genome is now available at the NCBI. The submitted genome and annotation comprised of completely annotated chromosome arms with associated genes and model transcripts contains 32,041 genes. No changes were made to the underlying *Arabidopsis* sequence for this release

Caenorhabditis elegans WS170

WormBase WS170 data freeze is currently available at the NCBI. As with the *Arabidopsis* release there are no changes to the underlying genomic sequence. WS170 contains 21,052 genes and their transcripts placed on the six *C. elegans* chromosomes. The gene, transcript, and protein annotation is provided by WormBase. NCBI provides gene predictions and calculates alignments to provide the *C. elegans* UniGene and transcript (EST) maps.

Access to Data

Sequences of the genome assemblies, transcripts, proteins, gene models, and gene records for the above genomes are available through the NCBI Entrez system. In addition, all sequences associated with these genomes can be searched using NCBI's Web BLAST services. Records are extensively integrated with other resources and databases in both Entrez and BLAST. In the Entrez system, annotated genes on these genomes are most effectively searched in the Entrez Gene database. For sequence data, a Genome BLAST page is available that allows searches against the genome and specialized sets of sequences including GenBank and RefSeq mRNA and protein sequences, expressed sequence tags, high throughput genomic sequence, whole genome shotgun, and trace archive sequences. Results of searches against the assembled genome can be displayed in the Map Viewer to provide essential genomic contextual information. Genomic BLAST may be reached via links on the Map Viewer home page (Figure 1). ¤



The Map Viewer provides a wide variety of genome mapping and sequencing data. More..

Q Search or Browse the Genome
B BLAST
G Genome Resources page

Annotation update plans for Human Genome build 36 Oct1
NCBI is planning to update its annotation of the human genom... more

Scientific name	Common name	Build	Tools
Ornithorhynchus anatinus	duck-billed platypus	Build 1.1	Q B G

Scientific name	Common name	Build	Tools
Monodelphis domestica	opossum	MonDom5	Q B G

Scientific name	Common name	Build	Tools
Bos taurus	cattle	Build 3.1 (based on Btau_3.1) Build 2.1 (based on Btau_2.0)	Q B Q B
Canis lupus familiaris	dog	Build 2.1	Q B

Figure 1. The new Map Viewer homepage with expandable organism sections providing links to the Maps (magnifying glass icon), genomic BLAST ("B" icon), and Genome Resources ("G" icon). Sections are expanded in this view to highlight some recent additions: duck-billed platypus, short-tailed opossum, and the marine phytoplankter, *Ostreococcus*.

BLAST® Lab

caaatccgttcttgatcgtacatagcgcatgtcagncaaatc
||||||| |||||||||||||| || |||||||| ||||||
caaatccattcttgatcgtacatggcacatgtcagtcaaatc

Web PSI-BLAST Now Produces Position-Specific Score Matrices

Position Specific Iterative BLAST (PSI-BLAST) is a popular tool that extends the sensitivity of standard protein blast. PSI-BLAST builds a position-specific scoring matrix (PSSM) from previous BLAST results and searches the database with this more sensitive position-specific scoring system. The ability to re-use PSI-BLAST PSSMs in searches against other databases is an important feature of both the standalone and the Web implementations of PSI-BLAST. Until recently, the encoded PSSMs produced by the Web PSI-BLAST were not compatible with the standalone version of PSI-BLAST, the blastpgp binary in the BLAST standalone package. Now, the new "PssmWithParameters" formatting option on the Web version produces the PSSM as a portable ASN.1 scoremat that can be downloaded and used in local searches with blastpgp or re-used on the Web service.

This edition of BLAST lab shows how to produce the new scoremat format in Web searches and re-use the PSSM in Web and standalone PSI-BLAST searches.

Producing the PSSM from a Web PSI-BLAST Search

A PSSM can be generated from the results of iteration two or higher of any Web PSI-BLAST search. From the Web results, clicking on the link to "Reformat these Results" brings up the formatting page with many options for redisplaying the results in various formats (Figure 1A). The "Show" pull-down list provides two options for obtaining the PSSM. Selecting the "PSSM" option and clicking the "View report" button will display the compressed ASCII encoded matrix in the Web browser (Figure 2 A). This can be saved as text and re-used in Web PSI-BLAST searches. However, the compressed ASCII encoded matrix cannot be used directly with standalone BLAST. Selecting the "PssmWithParameters" option and clicking the "View report" button will download the scoremat in the structured ASN.1 format to disk (Figure 2 B.) This scoremat is compatible with both the Web and standalone implementations of PSI-B LAST.

Using a saved PSSM in Web and Standalone BLAST

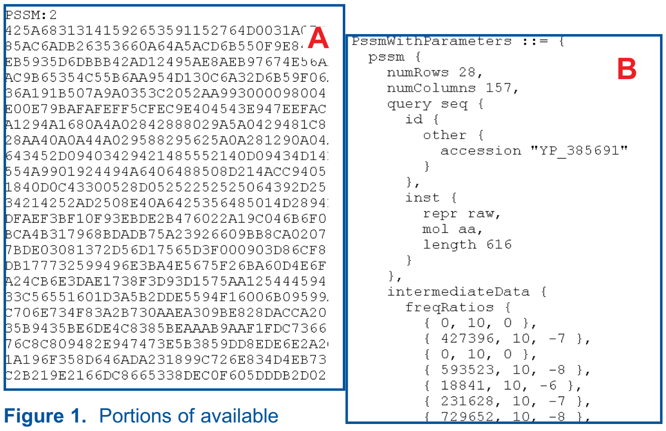
The saved ASCII encoded or scoremat PSSMs can be uploaded for use in other Web BLAST searches through the "Upload PSSM" option in the "Algorithm parameters " section of the PSI-BLAST Web form. Typically a PSSM created from one database is used in a search against a different one or additional rounds are performed to refine the PSSM. In all cases, the same query sequence used to generate the PSSM must be used in the subsequent search. The option for uploading the PSSM is found in the "Algorithm parameters" section of the PSI-BLAST form (Figure 2 B.) Figure 3 shows the results of a search with a third iteration PSSM generated from the non-redundant protein database. The results shown are from a search against the Reference Sequence protein database restricted to human sequences. The query is the conserved GAF domain-containing subsequence of a bacterial signaling protein from the *Geobacter metalloreducens* GS-15 genome (YP_385691, positions 205-361.) This search finds the corresponding GAF domain in human cGMP-binding phosphodiesterases. A standard blastp search using the same query sub-sequence does not find these matches in human proteins.

The scoremat PSSM saved from the Web search can also be used directly in standalone blastpgp for searching a local database. A typical command line is shown below. The query is the same sequence used to generate the PSSM from the Web; the -d argument takes the local formatted BLAST database name; the -R argument accepts the scoremat PSSM from the web search; and the -q 1 argument specifies that the scoremat is in ASCII text rather than binary form (-q 2).

```
blastpgp -i query  -d local_db -R web_pssm.asn -q 1 -o pgp_output.txt
```

The new ScoreMat PSSM format provides a powerful extension to NCBI Web PSI-BLAST searches that allows training a matrix on the large NCBI databases and using these sensitive matrices on custom local databases.

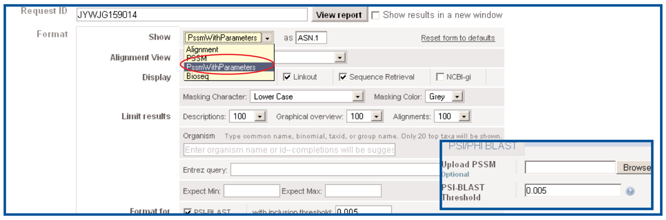
—TT



```
PSSM:2
425A6831314159265359115276 4D0031AC
85AC6ADB26353660A64A5ACD6B550F9E8
EB5935D6DBBB42AD12495AE8AEB97674E56A
AC9B65354C55B6AA954D130C6A32D6B59F06
36A191B507A9A0353C2052AA993000098004
E00E79BAFAFEFF5CFEC9E404543E947EEFAC
A1294A1680A4A02842888029A5A0429481C8
28AA40A0A44A029588295625A0A281290A04
643452D094034294214855521 40D09434D14
554A9901924494A6406488508D214ACC9405
1840D0C43300528D05252252525064392D25
34214252AD2508E40A6425356485014D2894
DFAEF3BF10F93EBDE2B476022A19C046B6F0
BCA4B317968BDADB75A23926609BB8CA0207
7BDE03081372D56D17565D3F000903D86CF8
DB177732599496E3BA4E5675F26BA60D4E6F
A24CB6E3DAE1738F3D93D1575AA125444594
33C56551601D3A5B2DDE5594F16006B09599
C706E734F83A2B730AAEA309BE828DACCA20
35B9435BE6DE4C8385BEAAAB9AAF1FDC7366
76C8C809482E947473E5B3859DD8EDE6E2A2
1A196F358D646ADA231899C726E834D4EB73
C2B219E2166DC8665338DEC0F605DDDB2D02
```

```
PssmWithParameters ::= {
  pssm {
    numRows 28,
    numColumns 157,
    query seq {
      id {
        other {
          accession "YP_385691"
        }
      },
      inst {
        repr raw,
        mol aa,
        length 616
      }
    },
    intermediateData {
      freqRatios {
        { 0, 10, 0 },
        { 427396, 10, -7 },
        { 0, 10, 0 },
        { 593523, 10, -8 },
        { 18841, 10, -6 },
        { 231628, 10, -7 },
        { 729652, 10, -8 },
```

Figure 1. Portions of available PSSM formats from the NCBI Web BLAST service. A. ASCII encoded PSSM compatibile only with Web service. B. Portable ASN.1 ScoreMat compatible with Web and standalone BLAST.



Request ID: JYWJG159014 | View report | □ Show results in a new window

Format

Show: PssmWithParameters as ASN.1

Alignment View

Display: ☑ Linkout ☑ Sequence Retrieval □ NCBI-gi

Masking Character: Lower Case   Masking Color: Grey

Limit results: Descriptions: 100   Graphical overview: 100   Alignments: 100

Organism: Type common name, binomial, taxid, or group name. Only 20 top taxa will be shown.

Enter organism name or id--completions will be sugges

Entrez query:

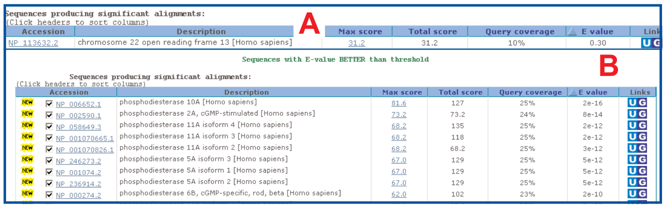
Expect Min: Expect Max:

Format for: ☑ PSI-BLAST with inclusion threshold: 0.005

Upload PSSM Optional

PSI-BLAST Threshold: 0.005

Figure 2. Controls on the Web BLAST forms for saving and uploading PSI-BLAST PSSMs. The Formatting page with available options for generating the ASCII encoded matrix (PSSM) or the ASN.1 ScoreMat (PssmWithParameters) is highlighted with a red circle. The Algorithm parameters section of the form for uploading a saved PSSM from disk is appears in the inset.



Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Links
NP_113632.2	chromosome 22 open reading frame 13 [Homo sapiens]	31.2	31.2	10%	0.30	UG

Sequences with E-value BETTER than threshold

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Links
NEW ☑ NP_006652.1	phosphodiesterase 10A [Homo sapiens]	81.6	127	25%	2e-16	UG
NEW ☑ NP_002590.1	phosphodiesterase 2A, cGMP-stimulated [Homo sapiens]	73.2	73.2	24%	8e-14	UG
NEW ☑ NP_058649.3	phosphodiesterase 11A isoform 4 [Homo sapiens]	68.2	135	25%	2e-12	UG
NEW ☑ NP_001070665.1	phosphodiesterase 11A isoform 3 [Homo sapiens]	68.2	118	25%	2e-12	UG
NEW ☑ NP_001070826.1	phosphodiesterase 11A isoform 2 [Homo sapiens]	68.2	68.2	25%	3e-12	UG
NEW ☑ NP_246273.2	phosphodiesterase 5A isoform 3 [Homo sapiens]	67.0	129	25%	5e-12	UG
NEW ☑ NP_001074.2	phosphodiesterase 5A isoform 1 [Homo sapiens]	67.0	129	25%	5e-12	UG
NEW ☑ NP_236914.2	phosphodiesterase 5A isoform 2 [Homo sapiens]	67.0	129	25%	5e-12	UG
NEW ☑ NP_000274.2	phosphodiesterase 6B, cGMP-specific, rod, beta [Homo sapiens]	62.0	102	23%	2e-10	UG

Figure 3. Protein BLAST searches using the GAF domain region of a signaling protein from the *Geobacter metalloreducens* genome (**YP_385691**, positions 205-361) against the Reference Protein database restricted to human proteins. A.) Ordinary protein BLAST results (RID: **JYZNJ1M9014**) showing no significant matches. B.) PSI-BLAST results (RID: **JYXE09V501R**) from human Reference Sequences using the Scoremat generated after three iterations of PSI-BLAST against the non-redundant protein database (nr) showing significant similarity to cGMP stimulated phosphodiesterases.

Department of Health and Human Services
Public Health Service, National Institutes of Health
National Library of Medicine
National Center for Biotechnology Information
Bldg. 38A, Room 3S308
8600 Rockville Pike
Bethesda, Maryland 20894